Optimizing jobs timeouts on clusters and production grids

Tristan Glatard[1,2], Johan Montagnat[1]
[1]*CNRS, I3S unit*
{*glatard,johan*}*@i3s.unice.fr*

Xavier Pennec[2]
[2]*INRIA Sophia-Antipolis, Asclepios*
Xavier.Pennec@sophia.inria.fr

Abstract

This paper presents a method to optimize the timeout value of computing jobs. It relies on a model of the job execution time that considers the job management system latency through a random variable. It also takes into account a proportion of outliers to model either reliable clusters or production grids characterized by faults causing jobs loss. Job management systems are first studied considering classical distributions. Different behaviors are exhibited, depending on the weight of the tail of the distribution and on the amount of outliers. Experimental results are then shown based on the latency distribution and outlier ratios measured on the EGEE grid infrastructure[1]. Those results show that using the optimal timeout value provided by our method reduces the impact of outliers and leads to a 1.36 speed-up even for reliable systems without outliers.

1. Introduction

A growing number of distributed applications is relying on large scale workload management systems. These applications usually trigger hundreds, thousands or even more jobs. Although large scale systems provide very high throughput as a direct consequence of the huge amount of resources available, they also introduce *high* and *variable* latencies that drastically impair the performances of the applications competing with other users' computations. The latency corresponds to the duration between the submission date and the one at which the job execution really starts. It includes the submission, scheduling and queuing times but also data transfers and delays in the monitoring system. Furthermore, less reliable systems such as production grids are also impacted by *outliers*.

The variability impacts jobs latency in a *normal operation mode*. It mainly comes from the heterogeneity of the infrastructure (endogen hardware and software factors) and from the load imposed to it (exogen factor). In this paper, we model the system latency by a probabilistic distribution.

[1] http://www.eu-egee.org/

Outliers correspond to *system faults* that lead to huge latencies prevailing on the ones of the other tasks of the application. Those latency values can be considered as infinite. Typical faults generating outliers are hardware failures, locally heavy load or scheduling errors leading to a job being queued in a extremely long queue. This outlier mode can be quantified by its proportion of jobs that never return.

Both variability and outliers penalize applications that rely on the completion of a high number of jobs. Indeed, a single job is then able to slow down the whole application. In case where there are dependencies between the jobs (*e.g.* in case of application workflows), the effects of variability and outliers are even more critical and lead to accumulated performance drops.

From the user point of view, strategies to reduce the impact of variability and outliers include multi-submission [1] (a given job is submitted many times and only the first completion is considered), jobs grouping (many jobs are grouped together to reduce the number of submissions) [2] and timeouting [3, 4]. Timeouting and resubmitting abnormally long jobs is a common strategy. Choosing the timeout value is often let to the administrator or the end user. However, a non trivial trade off has to be found as a too long timeout will penalize the jobs completion time too much, while a too short one may be overkilling, causing the unnecessary resubmission of jobs that almost completed. Hence, timeout strategies have been designed in areas as different as TCP throughput optimization [5], HTTP requests [6, 7] or power saving devices [8].

To determine an optimal timeout, we first model jobs execution times in section 2. We then present in section 3 some results of timeout optimization on classical distributions. To show how the optimization behaves on a real infrastructure, we are particularly interested in the asymptotic behavior of the system and on the impact of outliers. We finally present in section 4 some experimental results from a distribution of the latency measured on the EGEE production grid. Many of the proofs of the theoretical results are omitted in this paper and are detailed in a more detailed technical report [9].

R	grid latency
J	total time of a job (with resubmissions)
f_V/F_V	pdf/cdf of the random variable V
E_V	expectation of the random variable V
t_∞	timeout value
q	probability for a job to timeout
ρ	outliers ratio
ϕ/Φ	pdf/cdf of the standard normal distribution

Table 1. Notation table

2. Model of the job execution time

We adopted a probabilistic modeling of the large-scale workload manager. This approach has already been successfully reported to tackle related scheduling problems [10, 11]. We will denote random variables with capital letters whereas fixed values will be lowercase. For a random variable V, f_V denotes its probability density function (pdf) and F_V denotes its cumulative density function (cdf). Notations are summarized in table 1.

Let J be the total duration of a job (including all its potential resubmissions) and t_∞ be a user defined timeout value. The system is seen as a black box introducing a positive latency R on the job wall-clock time r in case of normal operation. The outlier ratio is denoted ρ. r is assumed to be a fixed value depending only on the job nature whereas R is a random variable.

We denote with q the probability for a job to timeout. A job timeouts either if it is an outlier or if it faces a latency which is superior to t_∞. Thus:

$$\begin{aligned} q &= \rho + (1-\rho)P(r+R>t_\infty) \\ \text{and then,}\quad q &= 1-(1-\rho)F_R(t_\infty - r). \end{aligned} \tag{1}$$

If the job timeouts, it is canceled and resubmitted. We neglect the cost of canceling and resubmitting a job as well as the resulting overload on the system, so that consecutive submissions are considered as independent. Let J_i be the duration of the job from its i^{th} submission to its completion. J_i can be recursively defined as:

$$J_i = \begin{cases} r+R & \text{with probability } 1-q \\ t_\infty + J_{i+1} & \text{with probability } q. \end{cases} \tag{2}$$

For the sake of clarity, we will assume that $r = 0$. This hypothesis is not restrictive. In the rest of the equations, it corresponds to the variable change $u = t - r$. In case of real job executions, r would have to be added to the timeout value. The goal is to express $J = J_1$ (through its cdf F_J) with respect to R and t_∞. J is superior to nt_∞ if and only if the job timed-out n times. Thus:

$$P(J > nt_\infty) = q^n \quad \text{so that} \quad P(J < nt_\infty) = 1-q^n. \tag{3}$$

We have the value of $F_J(x) = P(J < x)$ for every $x = nt_\infty$. We now have to obtain a complete expression of F_J. We can notice that $F_J(t)$ represents the probability for J to be inferior to t, so that **for all** t **in** $[nt_\infty, (n+1)t_\infty[$:

$$\begin{aligned} F_J(t) &= P(J < t | t \in [nt_\infty, (n+1)t_\infty]) \\ &= P(J < nt_\infty) + P(nt_\infty < J < t \mid t \le (n+1)t_\infty) \end{aligned}$$

and thus, according to equation 3:

$$F_J(t) = 1 - q^n + P(nt_\infty < J < t \mid t \le (n+1)t_\infty). \tag{4}$$

Given that $t \le (n+1)t_\infty$, a job duration J is in $[nt_\infty, t]$ if and only if the job timed-out n times (probability q^n) and succeeded on the $(n+1)^{\text{th}}$ attempt *i.e.* it was not an outlier (probability $1-\rho$) and $R \le t - nt_\infty$ (probability $F_R(t-nt_\infty)$). Therefore, $P(nt_\infty < J < t \mid t \le (n+1)t_\infty) = q^n(1-\rho)F_R(t - nt_\infty)$. We finally get, $\forall t \in [nt_\infty, (n+1)t_\infty[$:

$$F_J(t) = 1 - q^n + q^n(1-\rho)F_R(t - nt_\infty). \tag{5}$$

Given that $R>0$, $F_J(0) = 0$ and $\lim_\infty F_J = 1$. Moreover, F_J is continuous at every nt_∞. However, in general, F_J is not differentiable in $nt_\infty, \forall n$.

Note that if $\rho = 0$, then equation 5 resumes to:

$$F_J(t) = 1 - q^n + q^n F_R(t - nt_\infty) \quad \text{with} \quad q = 1 - F_R(t_\infty).$$

Thus, taking outliers into account corresponds to multiplying F_R by the $(1-\rho)$ factor.

2.1. Expectation of J

Computing the expectation of a job execution time, general conclusion can be made on its behavior when the timeout value increases, independently from the system latency distribution. The expectation of J is:

$$E_J(t_\infty) = \frac{1}{F_R(t_\infty)} \int_0^{t_\infty} u f_R(u) du + \frac{t_\infty}{(1-\rho)F_R(t_\infty)} - t_\infty. \tag{6}$$

Equation 6 compares to similar expressions derived for modeling completion times probabilistically : equation 6 in [4] and equation 1 in [3]. In both cases, the authors introduced a fixed cost penalty to resubmission that we consider to be zero (there is almost no overhead induced by job resubmission on a large scale system). In [4], the authors also derives higher moments of J and some relevant properties about them (*e.g.* their existence). Our hypotheses are similar to theirs except that they do not take into account outliers that are of major importance on the infrastructures we are targeting. As stated above, this parameter is characteristic of unreliable systems and they are needed to properly model a grid infrastructure. In [3], the authors do take into account the outlier ratio (denoted L). However, the studied

hypotheses do not really match ours : the case of a client being unable to hold more than one connection (so-called simple client) is not developed, even if noticeable remarks (such as the fact that the timeout values of all the resubmissions have to be identical) are done.

E_J has the following limits:

$$\lim_{t_\infty \to \infty} E_J(t_\infty) = +\infty \quad \text{if } \rho \neq 0 \tag{7}$$

$$\text{and} \quad \lim_{t_\infty \to \infty} E_J(t_\infty) = E_R \quad \text{otherwise.} \tag{8}$$

If $\rho \neq 0$, the line $E_R + \frac{\rho}{1-\rho} t_\infty$ is an asymptote of $E_J(t_\infty)$.

The first limit can be explained by noticing that a single outlier may lead to an infinite execution time. When $t_\infty \to +\infty$, the probability for encountering an outlier tends towards 1 and the expected execution time tends towards infinity. The second limit is also intuitive: in absence of outliers, if no timeout value is set, then the system latency would not be disturbed and the expectation of a job duration would resume to the one of the system latency.

Equations 7 and 8 show that ρ has a major impact on the system behavior. The case $\rho = 0$ corresponds to a reliable cluster management system: faults causing jobs loss are very unlikely (highly reliable LAN, robust schedulers). The case $\rho > 0$ is needed to model grid infrastructures. Lower reliability of WANs, scale effects and less mature workload management middlewares lead to a significant number of outliers. For instance on the EGEE infrastructure, ρ is in the order of 2%. In case of outliers it is mandatory to set a timeout value.

3. Results on classical distributions

In this section, we study some classical distributions from a theoretical point of view in order to understand how the timeout value impacts the expectation of the job duration both with and without outliers. We explore distributions with light tails (uniform and truncated Gaussian), heavy tails (log-normal) and power tails (Pareto) to show how they exhibit different behaviors. The exponential distribution will constitute a transition between light and heavy-tailed distributions. Light-tailed distributions are the ones that decay faster than the exponential. In this case, there exists a such that: $\lim_{t\to+\infty} e^{at}(1 - F(t)) = 0$. On the contrary, heavy-tailed distributions decay more slowly than the exponential : $\lim_{t\to+\infty} e^{at}(1 - F(t)) = +\infty$. Power-tailed distributions are a subset of the heavy-tailed ones. In this case, there exists a and b such that $\lim_{t\to+\infty} \frac{1-F(t)}{t^a} = b$.

For each distribution, our goal is to determine the optimal timeout value $\hat{t}_\infty = \arg\min_{t_\infty}\{E_J(t_\infty)\}$. In case of very reliable systems (when no outliers are present), the optimal value of the timeout may be $+\infty$, which means that no timeout should be set. Another singular optimal timeout value is 0. This configuration occurs when the probability for the job to face a null latency is so high that it is interesting to resubmit the job as soon as one knows that it is going to face a non null latency. This later result would only be realistic if it was possible to resubmit an arbitrarily large number of jobs at no additional cost. Obviously, the overhead induced on any real system would finally slow down the process.

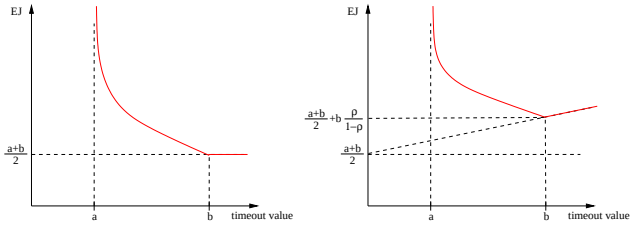

Figure 1. Behavior of $E_J(t_\infty)$ for the uniform distribution without (left) and with (right) outliers.

3.1. Uniform distribution

In this case, the pdf of the system latency is:

$$f_R(t) = \begin{cases} \frac{1}{b-a} & \text{if } t \in [a,b] \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

We can derive from equation 6 the expectation of J:

$$E_J(t_\infty) = \begin{cases} +\infty & \text{if } t_\infty \leq a \\ \frac{t_\infty + a}{2} + t_\infty \frac{b - t_\infty + \rho(t_\infty - a)}{(t_\infty - a)(1-\rho)} & \text{if } t_\infty \in [a,b] \\ \frac{b+a}{2} + t_\infty \frac{\rho}{1-\rho} & \text{otherwise.} \end{cases} \tag{10}$$

The curve of $E_J(t_\infty)$ is depicted on figure 1. The optimal timeout value is b both with and without outliers. Without outliers, setting the timeout to $+\infty$ is also optimal because the expectation of J is constant in [b,+∞[.

3.2. Truncated Gaussian

Normal distributions are the most commonly used but they do not exclude negative values. In our case, the latency cannot be lower than 0. We are thus considering Gaussian distributions with mean μ and standard-deviation σ truncated above 0. The pdf and cdf of the system latency are:

$$f_R(t) = \begin{cases} \frac{1}{\Phi\left(\frac{\mu}{\sigma}\right)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} & \text{if} \quad t \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$F_R(t) = \frac{\Phi\left(\frac{\mu}{\sigma}\right) - \Phi\left(\frac{\mu - t}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right)} \quad \text{with} \quad \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}u^2} du.$$

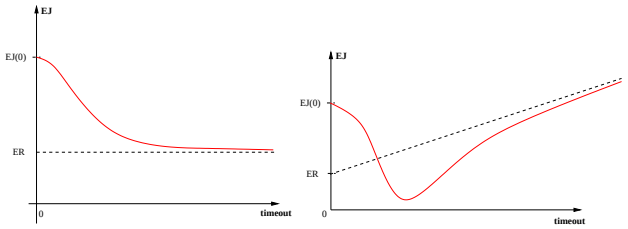

Figure 2. Behavior of the expectation of J for a truncated Gaussian distribution without (left) and with (right) outliers.

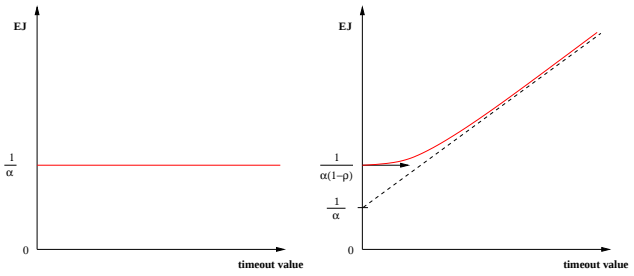

Figure 3. Behavior of E_J for an exponential distribution without (left) and with (right) outliers.

The expectation of the job duration is then:

$$\mu + \sigma \frac{\phi\left(\frac{\mu}{\sigma}\right) - \phi\left(\frac{\mu - t_\infty}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right) - \Phi\left(\frac{\mu - t_\infty}{\sigma}\right)} + \frac{1}{1-\rho} t_\infty \left(\frac{\Phi\left(\frac{\mu - t_\infty}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right) - \Phi\left(\frac{\mu - t_\infty}{\sigma}\right)} + \rho \right).$$

with $\phi = \Phi'$.

The curve of E_J is plotted on figure 2. E_J exhibits different behaviors depending on the presence of outliers or not. If there are no outliers ($\rho = 0$), E_J is decreasing towards its limit E_R when $t_\infty \to +\infty$. On the other hand, when $\rho \neq 0$, then E_J exhibits a global minimum reached for $\hat{t}_\infty < +\infty$. The corresponding proof is based on the fact that the forth derivative of E_J is always positive, so that we can study the existence of a root in the lower order derivatives.

If the distribution of the system latency is Gaussian and $\rho = 0$, timeouting is not a solution to limit the impact of variability, regardless of the variability order of magnitude. In this case, other solutions such as multi-submissions or job grouping have to be studied.

3.3. Exponential distribution

In this case, the cdf of the system latency is:

$$F_R(t) = 1 - e^{-\alpha t}.$$

And according to equation 6, the expectation of J is:

$$E_J(t_\infty) = \frac{1}{\alpha} + \frac{\rho t_\infty}{(1-\rho)\left(1 - e^{-\alpha t_\infty}\right)}.$$

The curve of $E_J(t_\infty)$ is depicted on figure 3. In case of outliers, E_J is increasing and the best timeout value is $\hat{t}_\infty = 0$. If there are no outliers, the expectation of J is independent from t_∞, which is a singular behavior particular to the exponential distribution. This characteristic of the exponential distribution has to be related to the fact that this distribution is the only one to be memory-less. In this case, at a given instant, knowing that a job is still in the system does not give any information about its future behavior and the timeout value thus does not impact the distribution of J.

3.4. Log-normal distribution

The log-normal distribution is a typical example of heavy-tailed distribution. In some cases, it is able to fit job runtimes on clusters [12]. In this section, we assume that R has a log-normal distribution with parameters μ and σ. In this case, the cdf and pdf of the system latency are:

$$F_R(t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right) \quad \text{and} \quad f_R(t) = \frac{1}{t\sqrt{2\pi}\sigma} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}}.$$

The expectation and standard-deviation of R are:

$$E_R = e^{\mu + \frac{\sigma^2}{2}} \quad \text{and} \quad \sigma_R = \left(e^{\sigma^2} - 1\right) e^{2\mu + \sigma^2}. \tag{11}$$

In this case, we can show that E_J is:

$$E_R.\left(\frac{\Phi\left(x_\infty - \sigma\right)}{\Phi(x_\infty)} + e^{\sigma x_\infty - \frac{\sigma^2}{2}} \left(\frac{1}{(1-\rho)\Phi(x_\infty)} - 1 \right) \right) \tag{12}$$

$$\text{where} \quad x_\infty = \frac{\ln(t_\infty) - \mu}{\sigma}$$

This expression shows that the minimization of E_J can be performed independently from μ on the transformed variable x_∞. The obtained solution $\hat{x}_\infty(\sigma, \rho)$ only depends on σ and ρ. The optimal timeout value can then be written as:

$$\hat{t}_\infty(\mu, \sigma) = e^{\mu} K(\sigma, \rho) \quad \text{where} \quad K(\sigma, \rho) = e^{\sigma \hat{x}_\infty(\sigma, \rho)} \tag{13}$$

and:

$$\hat{x}_\infty(\sigma, \rho) = \arg\min_{x_\infty} \left(\frac{\Phi\left(x_\infty - \sigma\right)}{\Phi(x_\infty)} + e^{\sigma x_\infty - \frac{\sigma^2}{2}} \left(\frac{1}{(1-\rho)\Phi(x_\infty)} - 1 \right) \right).$$

$K(\sigma, \rho)$ is actually the optimal timeout value for $\mu = 0$.

We also have the following limit for $t_\infty = 0$:

$$\lim_{t_\infty \to 0} E_J(t_\infty) = \lim_{x_\infty \to -\infty} E_J(x_\infty) = +\infty.$$

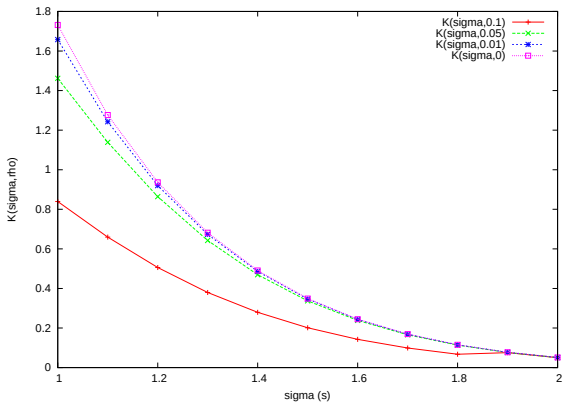

Figure 4. Evolution of the optimal timeout value for μ=0 in the log-normal case.

This infinite limit proves that when $\rho \neq 0$, there exists a finite non null optimal timeout value that minimizes E_J. Indeed, in this case, the limit of E_J when t_∞ tends towards infinity is infinite, according to equation 7 and E_J thus has to reach a global minimum.

The existence of a global minimum of $E_J(t_\infty)$ when $\rho \neq 0$ is not straight-forward. Given the infinite limit of E_J when t_∞ tends towards 0 and given that $E_J(+\infty) = E_R$, it resumes to the existence of a t_∞ for which $E_J(t_\infty)<E_R$. If $\sigma > 1$, then $t_\infty = e^\mu$ satisfies this relation. Indeed, in this case, $x_\infty = 0$ and according to equation 12, $E_J(x_\infty = 0) = E_R(2\Phi(-\sigma) + e^{-\frac{\sigma^2}{2}})$. A numeric resolution then shows that $E_J<E_R$ if and only if $\sigma \gtrsim 0.9311$. Numeric simulations suggest that E_J has a global minimum even for lower values of σ. However, an analytic proof still has to be done.

Figure 4 displays a simulation of the optimal timeout value for μ=0, several values of the outlier ratio and σ ranging from 1 to 2. We first can notice that $K(\sigma, \rho)$ seems decreasing with respect to ρ. The timeout value thus has to be reduced when the proportion of outliers is increasing, which is coherent. Moreover, given an outlier ratio, the optimal timeout value for $\mu = 0$ is decreasing as σ is growing. It is also coherent because the standard-deviation of the log-normal distribution is increasing with respect to σ (see equation 11). The optimal timeout value thus has to be reduced as the variability of the infrastructure is growing.

3.5. Pareto distribution

The Pareto distribution was introduced to represent the distribution of wealth and proved to be very accurate to model a large class of computer systems measurements (jobs durations, size of the files, data transfers length on the Internet...) [13]. It is an example of power tailed distribution. The cdf of the system latency is then:

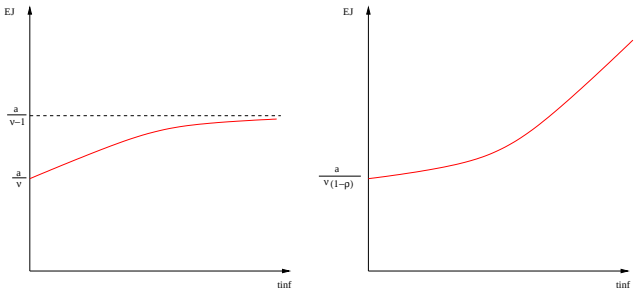

Figure 5. Behavior of E_J for a Pareto distribution of R. Left: no outliers ; Right: $\rho \neq 0$.

$$F_R(t) = 1 - \left(\frac{a}{a+t}\right)^\nu \quad \text{with} \quad a \quad \text{and} \quad \nu > 0.$$

The expectation is only defined for ν>1. Then:

$$E_R = \frac{a}{\nu - 1}.$$

In this case, the expression of E_J can be directly derived from equation 6 and it is:

$$\frac{a + t_\infty \nu - a\left(\frac{a+t_\infty}{a}\right)^\nu}{(1-\nu)\left[\left(\frac{a+t_\infty}{a}\right)^\nu - 1\right]} + \frac{t_\infty}{(1-\rho)\left[\left(\frac{a+t_\infty}{a}\right)^\nu - 1\right]} + \frac{\rho}{1-\rho} t_\infty.$$

We also have the following limit when the timeout value is null:

$$\lim_{t_\infty \to 0} E_J(t_\infty) = \frac{a}{\nu(1-\rho)}.$$

We then can show that the expectation of the job duration time is increasing with respect to the timeout value, regardless of the ρ value. The optimal timeout value is thus 0. The behavior of $E_J(t_\infty)$ is depicted on figure 5.

3.6. Results summary and interpretation

Table 2 displays a summary of the results we obtained for various distributions of the system latency. Those results suggest that the weight of the tail of the distribution of the system latency is a discriminatory parameter for the timeout optimization when outliers are not present. Indeed, only heavy-tailed distributions such as the log-normal, or the Pareto ones lead to finite optimal timeout values. In this case, which corresponds to the most realistic one [12], the optimization speeds the execution up. On the other hand, when the distribution of the system latency decays more rapidly than the exponential, then setting a timeout value always penalizes the execution and the optimal timeout is

Distribution of R	Without outliers ($\rho = 0$)	With outliers (ρ>0)
Uniform	$+\infty$ (or b)	b
Trunc. Gaussian	$+\infty$	$0 < \hat{t}_\infty < +\infty$
Exponential	any	0
Log-normal (μ,σ)	$\hat{t}_\infty = e^{\mu}K(\sigma) < +\infty$	$0 < \hat{t}_\infty < +\infty$
Pareto (ν>1)	0	0

Table 2. Optimal timeout values

$+\infty$. The exponential distribution stands in the middle and is not affected by the timeout value.

As noticed in section 2, taking into account the outliers resumes to multiplying F_R by the factor (1-ρ). In this case, the distribution of the system latency becomes heavy-tailed as $\lim_{x\to+\infty} e^{ax}(1-(1-\rho)F_R(x)) = +\infty$ when $a > 0$. Indeed, in this case the optimal timeout value that we found is always finite, which is coherent with this interpretation.

3.7. Performance improvement

In case of reliable systems (without outliers), the expectation of the job duration without timeout equals to the one of the system latency. In this case, the ratio $\frac{E_R}{E_J(\hat{t}_\infty)}$ evaluates the speed-up yielded by the optimization. If the latency of the system is light-tailed, then setting a timeout value always penalizes the execution. The best strategy is thus to set the timeout value to infinity. In this case, the optimization does not provide any speed-up with respect to the expectation of the system latency. Concerning the limit case of an exponential distribution, the expectation of the job duration is independent from the timeout value and the optimization does not lead to any speed-up.

The optimization becomes interesting for heavy-tailed distributions as already suggested. For the log-normal case, figure 6 displays a numerical simulation of the evolution of the speed-up of the optimization with respect to σ for a particular value of μ showing that the speed-up is growing with σ. In this case, both E_R and σ_R are also growing with σ (see equation 11). Thus, the higher and the more variable the latency, the more interesting the timeout optimization.

Concerning the Pareto distribution, the optimized expectation of the job duration without outliers is $\frac{a}{\nu}$, whereas the one obtained without setting any timeout is $E_R = \frac{a}{\nu-1}$. The speed-up obtained by the optimization is thus $\frac{\nu}{\nu-1}$. This value is maximal for $\nu = 1$ and decreases towards 1 when ν increases. Under Pareto assumption, the variance of the system latency $\left(\frac{\nu a^2}{(\nu-2)(\nu-1)^2}\right)$ is decreasing with respect to ν. Here again, the more variable the latency of the infrastructure, the higher the speed-up yielded by the optimization.

When outliers are present, the optimization of the timeout prevents the expectation of J to be infinite. The impact of the optimization can then be evaluated by comparing the optimized expectation of the job duration to the one obtained without outliers. In case of a uniform distribution, outliers add the term $b\frac{\rho}{1-\rho}$ to the expectation of the job duration. This term is increasing with respect to the outlier ratio and tends towards infinity when ρ tends towards 1. The exponential distribution and the Pareto one exhibit a similar behavior: the outliers introduce an extra $\frac{1}{1-\rho}$ factor on the expectation of the job duration.

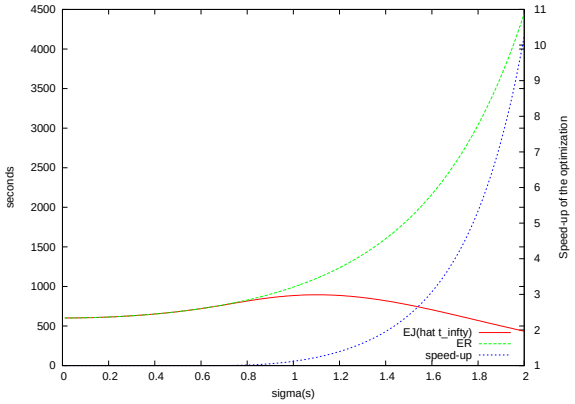

Figure 6. Evolution of the speed-up of the optimization for μ = 6.4s in the log-normal case.

4. Experiments

In this section, we present experimental results obtained by measuring the distribution of the latency of the EGEE grid infrastructure on a particular time period. The EGEE grid is a pool of thousands computers (standard PCs) and storage resources accessible through the gLite middleware. The resources are operated in computing centers, each of them running its internal batch scheduler. Jobs are submitted from a user interface to a central Resource Broker (RB) which dispatches them to the computing centers. EGEE is a production infrastructure with more than 25000 CPUs spread in more than 190 computing centers. It is characterized by its high throughput but also by its high latency, high variability and outliers. It is thus an ideal target to test our optimization procedure.

4.1. Measure of the distribution

To measure the distribution of the system latency on the EGEE grid, we submitted probe jobs that only consist in the execution of a `/bin/hostname` and we measure their round-trip time. We maintain a constant number of probes inside the system by submitting a new one as soon as one completed to avoid introducing any extra variability.

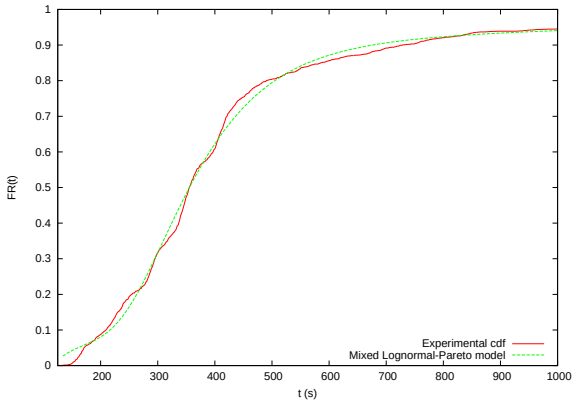

Figure 7. Measured data (plain) and best fitting Log-normal-Pareto model (dashed).

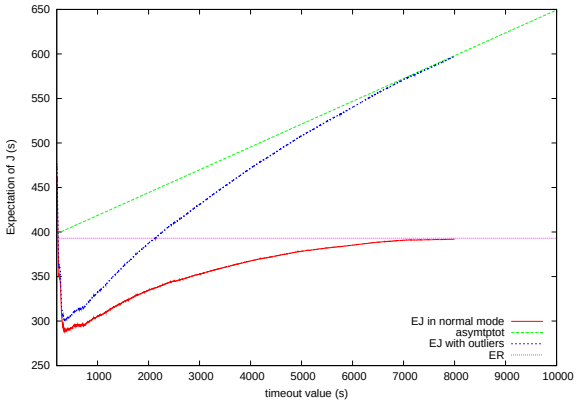

Figure 8. Evolution of the expectation of J in the experimental case.

Our measure of the distribution of R gathers 2137 probe jobs spread over 3 different days and involving 3 RBs. The maximal duration of those jobs was fixed to $t_{\max} = 10000$ seconds. Beyond this value, we consider a job as an outlier. Given those conditions, we obtained an outlier ratio of 2.5%. In normal operating mode, the measured distribution of R is plotted on figure 7 (plain red curve). Its expectation is 393 seconds and its standard deviation is 792 seconds.

4.2. Timeout optimization

If we do not take outliers into account, the evolution of E_J with respect to the timeout value is plotted on figure 8 (plain red curve). E_J then converges towards E_R as predicted by the theoretical analysis. In this case, E_J reaches a minimum for $\hat{t}_\infty = 360$s. At this optimal point, $\hat{E}_J(\hat{t}_\infty) = 289$s. The speed-up w.r.t to an execution without timeout is 1.36.

The evolution of $E_J(t_\infty)$ taking the outliers into account is plotted on figure 8 (dashed green curve). E_J effectively tends to its asymptote. The optimal timeout value $\hat{t}_\infty$ is now 358 seconds and $\hat{E}_J(\hat{t}_\infty)$ has grown to 300 seconds. Setting the optimal timeout value thus limits the impact of the outliers to a 11-seconds loss, whereas it would be highly superior if the timeout value is not properly set, as suggested by figure 8.

This figure also shows that the timeout value should better be overestimated than underestimated. Both curves are rapidly decreasing to the optimal timeout value whereas they increase more smoothly after it.

4.3. Model of the measured distribution

To relate those experimental results to the ones presented in the previous section, we here model the experimental distribution of the grid latency. The experimental data shown on figure 7 cannot be reasonably fitted with any of the standard distributions described in section 3. However, the distribution appears to be close to a log-normal distribution for low values (up to 500 seconds) and a Pareto distribution beyond. Based on this observation we fitted the experimental data with the following distribution which is an interpolation of the log-normal and Pareto ones, for t in $[t_{\min}, t_{\max}]$:

$$F_R^m(t) = (1-\alpha(t))\,\Phi\left(\frac{\ln(t-t_{\min})-\mu}{\sigma}\right) + \alpha(t)\left(1-\left(\frac{a}{a+t}\right)^{\nu}\right) \quad \text{with} \quad \alpha(t) = \left(\frac{t-t_{\min}}{t_{\max}-t_{\min}}\right)^{k}. \tag{14}$$

$t_{\min}$ denotes the smallest latency measured among the data (the cdf is zero below this value) and $t_{\max}$ the highest one. There are thus five parameters fully describing this model (μ, σ, a, ν, and k). $\alpha(t)$ is a weight function designed so that $\alpha(t_{\min}) = 0$ and $\alpha(t_{\max}) = 1$. The model thus tends to a log-normal distribution in $t_{\min}$ and to a Pareto one in $t_{\max}$.

We have estimated the best fit of the model 14 with the experimental data by least-square minimization. The optimal model is displayed on figure 7 (dashed green curve). A Kolmogorov-Smirnov test was made to evaluate the quality of the model. When considering an undersampling of up to 1000 measurements, the Kolmogorov-Smirnov test value is $D_{1000} = 1.35$ (we used $D_n = \sqrt{n}\sup|F_R^{exp} - F_R^m|$), which correspond to a p-value $p = 0.051$. The tests is thus positive. It shows that a simple model (5 parameters) can accurately model the distribution measured over a very complex workload system (EGEE grid infrastructure) even when considering a very large data sample.

4.4. Practical use of the method

Exploiting our method in practice requires (i) the estimation of the distribution of the grid latency and (ii) the deriva-

tion of the optimal timeout value. The first issue cannot reasonably be addressed by an end-user. Collecting live statistics about the grid latency implies the submission of several probe jobs that may disturb the grid functioning without any production usage of the resources. However, such an information should easily be available from logs of the grid workload management system. A production grid such as EGEE already includes a logging service which would be able to compute and update the cdf of the grid latency over time. Yet, more fundamental problems such as the handling of non-stationarities of the workload still remains and are addressed by works such as [14].

Once the distribution of the grid latency is available, deriving the optimal timeout value with our method is easily automatable. The optimization criterion (*i.e.* the expectation of the job execution time written in equation 6) is rapidly computable: it mainly includes the computation of an integral of $u f_R(u)$, which is a piecewise linear function when an empirical distribution is considered. For instance, plotting the curves of figure 8 takes less than 2 seconds on a modern PC. The optimization procedure itself is monodimensional and easily implementable too.

5. Conclusion

We presented a probabilistic model of jobs execution time taking into account timeouting and resubmissions. It can describe both job management systems prone to face outliers (grid) or not (cluster). The optimal timeout value highly depends on the distribution of the system latency. Without outliers, the heavy-tailed distributions lead to a finite optimal timeout value whereas for the light-tailed ones setting a timeout value always penalizes the execution.

When considering outliers, our model predicts that the job execution time w.r.t the timeout value is diverging to $+\infty$ for every distributions. The expectation of J has an asymptote whose slope only depends on the outlier ratio. The optimal timeout value is then finite for all the studied distributions since taking outliers into account lengthens the tail of the distribution.

We finally presented some results from an empirical distribution from the EGEE grid. It is heavy-tailed and modelizable through a mixture of log-normal and Pareto distributions. Even without outliers, a 1.36 speed-up can be achieved by optimizing the timeout value. Considering outliers, optimizing the timeout value is even more critical and the resulting expectation of the job duration is close to the one obtained without outliers.

6. Acknowledgments

We thank Pierre Bernhard (University of Nice Sophia-Antipolis) and Philippe Nain (INRIA Sophia-Antipolis) for their useful advices. This work is partially funded by the French program "ACI-Masse de données", AGIR project. We are grateful to the EGEE European project for providing the grid infrastructure and user assistance.

References

[1] H. Casanova, "On the Harmfulness of Redundant Batch Requests," in *15th IEEE International Symposium on High Performance Distributed Computing (HPDC'06)*, (Paris, France), pp. 255–266, June 2006.

[2] T. Glatard, J. Montagnat, and X. Pennec, "Probabilistic and dynamic optimization of job partitioning on a grid infrastructure," in *Parallel, Distributed and network-based Processing*, (Montbéliard-Sochaux, France), pp. 231–238, 2006.

[3] L. Libman and A. Orda, "Optimal Retrial and Timeout Strategies for Accessing Network Resources," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 551–564, 2002.

[4] A. van Moorsel and K. Wolter, "Analysis of Restart Mechanisms in Software Systems," *IEEE Trans. on Software Eng.*, vol. 32, pp. 547–558, Aug. 2006.

[5] A. Kesselman and Y. Mansour, "Optimizing TCP Retransmission Timeout," in *International Conference on Networking*, vol. 3421 of *LNCS*, (Saint-Denis de la Réunion), 2005.

[6] P. Reinecke, A. van Moorsel, and K. Wolter, "A Measurement Study of the Interplay between Application Level Restart and Transport Protocol," in *Intl Service Availability Symposium*, vol. 3335 of *LNCS*, (Munich, Germany), pp. 86–100, May 2004.

[7] W. Xie, H. Sun, Y. Cao, and K. Trivedi, "Optimal Webserver Session Timeout Settings for Web Users," in *Computer Measurement Group Conferenceries*, (Reno, NV, USA), pp. 799–820, Dec. 2002.

[8] P. Rong and M. Pedram, "Determining the optimal timeout values for a power-managed system based on the theory of Markovian processes: offline and online algorithms," in *Design, Automation and Test in Europe*, (Munich, Germany), pp. 1128–1133, Mar. 2006.

[9] T. Glatard, J. Montagnat, and X. Pennec, "Optimizing jobs timeouts on clusters and production grids," tech. rep., Laboratoire Images Signaux Systèmes, Sophia-Antipolis, France, Nov. 2006.

[10] H. Gautama and A. J. C. van Gemund, "Symbolic Performance Estimation Of Speculative Parallel Programs," *Parallel Processing Letters*, vol. 13, no. 4, pp. 513–524, 2003.

[11] J. Schopf and F. Berman, "Stochastic Scheduling," in *Supercomputing*, (Portland, USA), 1999.

[12] H. Li, D. Groep, and L. Wolters, "Workload Characteristics of a Multi-cluster Supercomputer," in *Job Scheduling Strategies for Parallel Processing*, pp. 176–193, 2004.

[13] M. Harchol-Balter, "Task Assignment with Unknown Duration," *Journal of the ACM*, vol. 49, pp. 260–288, Mar. 2002.

[14] D. Feitelson, *Workload modeling for performance evaluation*, pp. 114–141. LNCS vol 2459, Sept. 2002.